BGI Sequencing Data Report

2023/5/15

Table of Contents

1 Project Information

Project code: F20FTSEUHT0946-02_MUSbeusR_poolC
Sample number: 12

2 Data Statistics

Raw reads produced from sequencer contain adapters, unknown or low quality bases. The statistics of raw data is shown below.

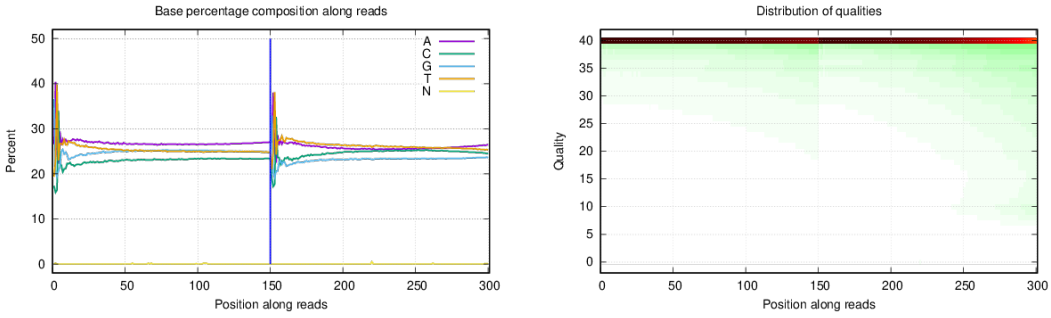
Sample	Length	Q20(%)	Q30(%)	GC Content(%)	Total Reads	Total Bases
BD1_unc	150;150	97.93;95.46	94.02;89.10	47.81;47.77	23,305,722	6,991,716,600
CB1_unc	150;150	97.49;94.24	92.93;86.38	49.30;49.30	42,114,499	12,634,349,700
CB1_wt	150;150	97.58;94.46	93.11;86.81	48.63;48.62	46,843,361	14,053,008,300
CB2_unc	150;150	97.90;96.34	94.31;90.99	46.65;46.12	667	200,100
CS	150;150	100.00;98.00	99.33;94.00	50.00;54.33	2	600
Liv_hic_unc_01	150;150	98.00;94.93	94.08;87.94	47.48;47.47	2,917	875,100
Liv_hic_unc_03	150;150	97.90;95.55	93.68;88.90	46.17;46.11	23,314,924	6,994,477,200
Liv_hic_wt_01	150;150	97.72;95.02	93.12;87.81	46.60;46.33	227	68,100
Liv_hic_wt_03	150;150	97.90;96.04	93.68;89.96	45.24;45.19	17,900,492	5,370,147,600
kid_unc3	150;150	97.93;95.53	93.94;89.11	47.72;47.70	28,109,032	8,432,709,600
kid_wt3	150;150	97.85;95.53	93.65;89.05	46.64;46.63	39,516,587	11,854,976,100
kid_wt4	150;150	97.97;95.56	93.96;88.97	46.76;46.70	21,512,951	6,453,885,300

Table Format:
1. Sample: The name of sample
2. Length: The Length of reads
3. Q20 (%): The proportion of nucleotides with quality value larger than 20
4. Q30 (%): The proportion of nucleotides with quality value larger than 30
4. GC Content(%): The proportion of bases G and C
5. Total Reads: The total number of raw read pairs
6. Total Bases: The total nucleotides number of raw reads

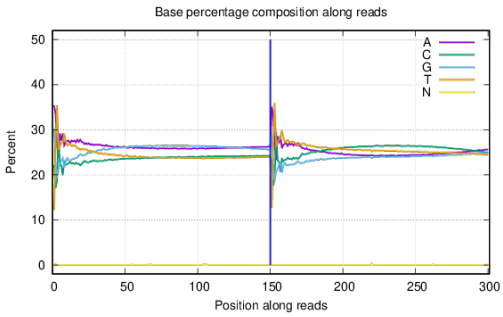
3 Data Quality Control

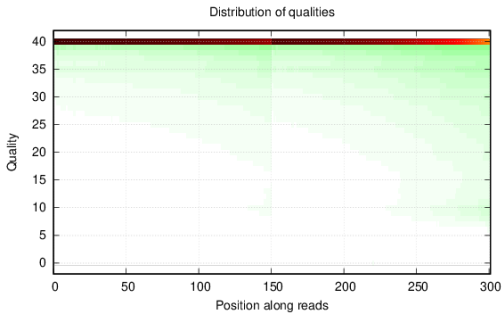
The distribution of base percentage and qualities along reads in data filtering are shown as following(If a sample has multiple lanes, only one of them will be displayed). The left picture is base percentage distribution along reads the sample, the right picture is distribution of qualities along reads of the sample.

Quality control of sample BD1_unc

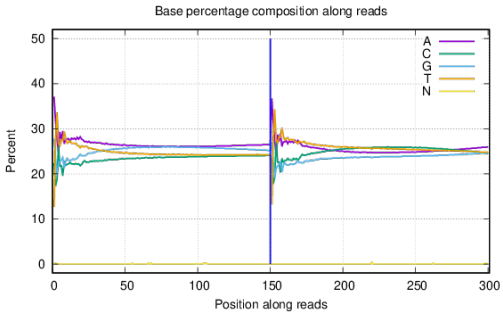


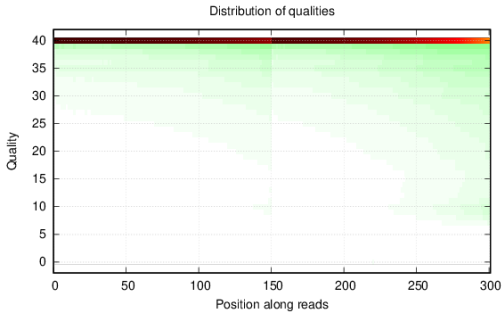
Quality control of sample CB1_unc



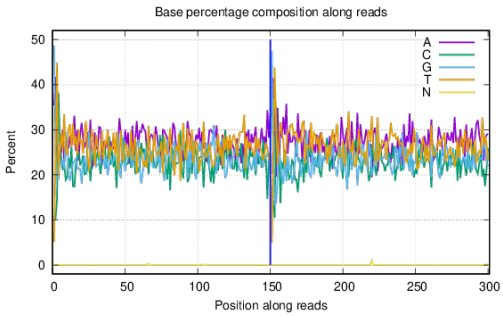


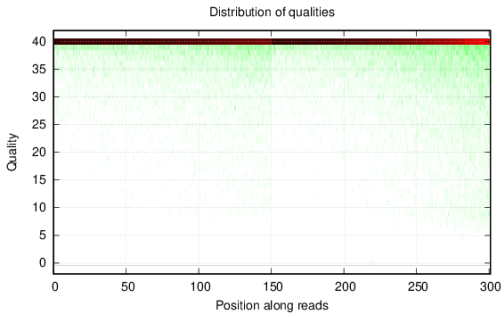
Quality control of sample CB1_wt



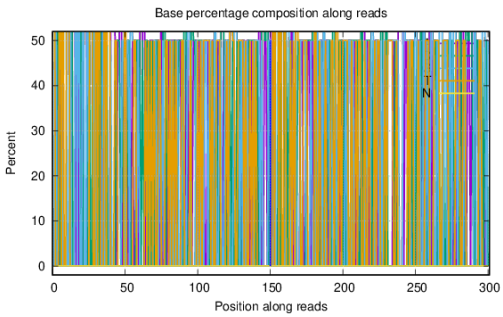


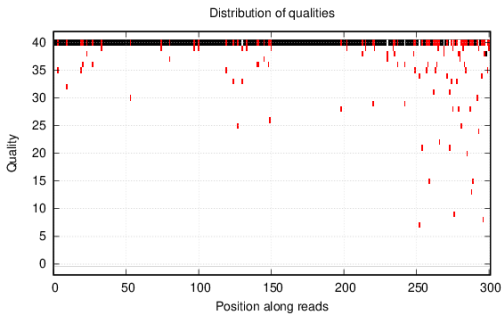
Quality control of sample CB2_unc



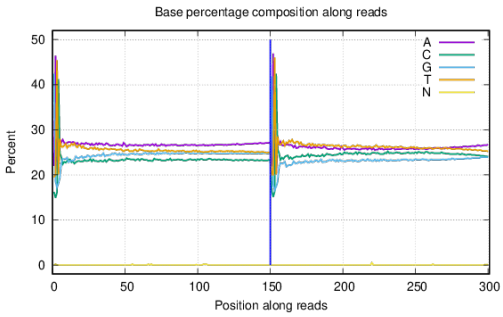


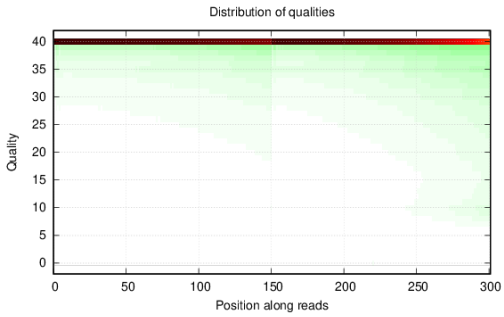
Quality control of sample CS



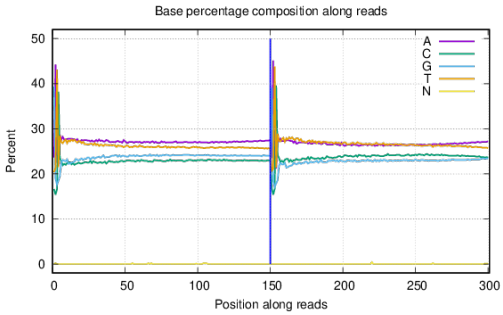


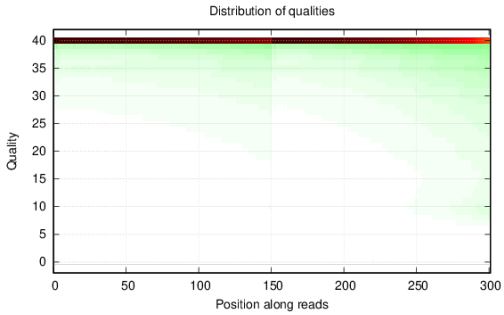
Quality control of sample kid_unc3



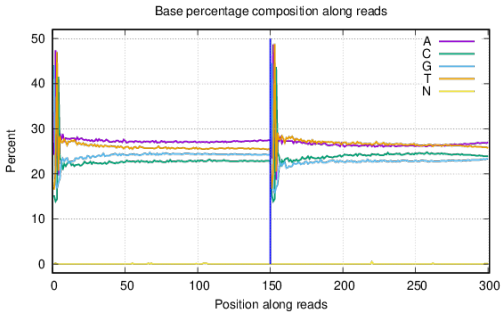


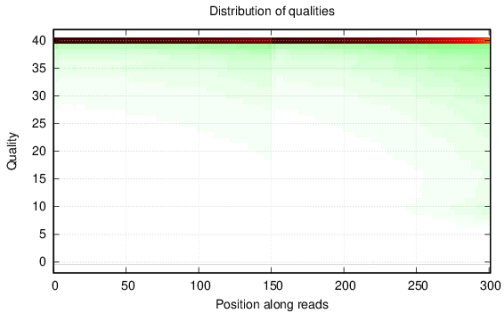
Quality control of sample kid_wt3



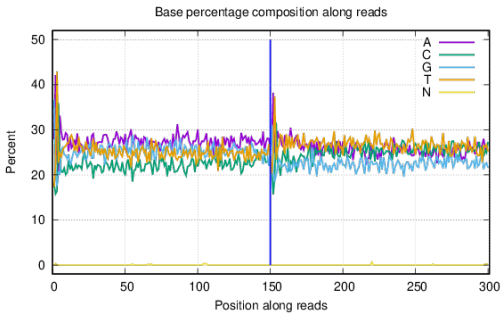


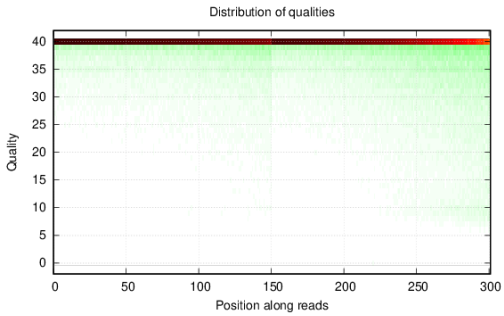
Quality control of sample kid_wt4



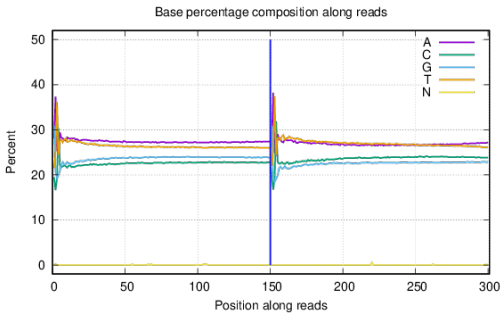


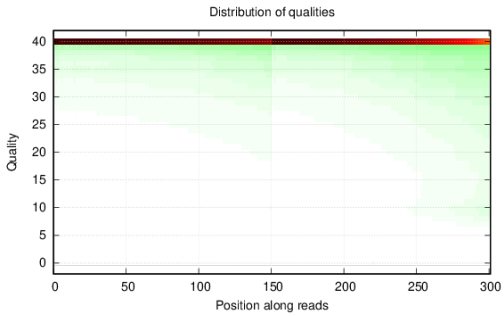
Quality control of sample Liv_hic_unc_01



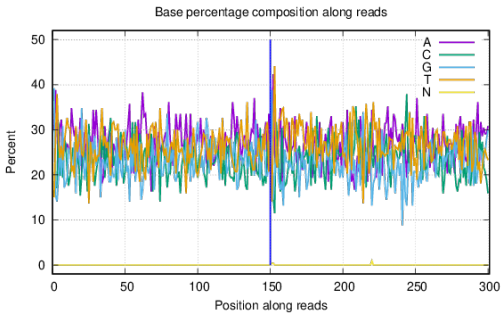


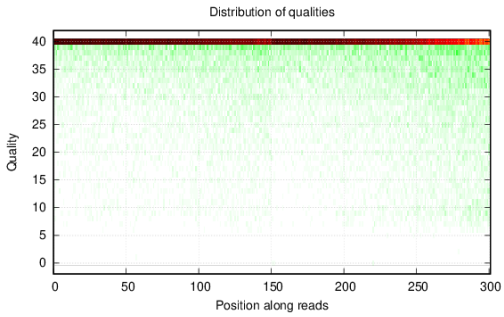
Quality control of sample Liv_hic_unc_03



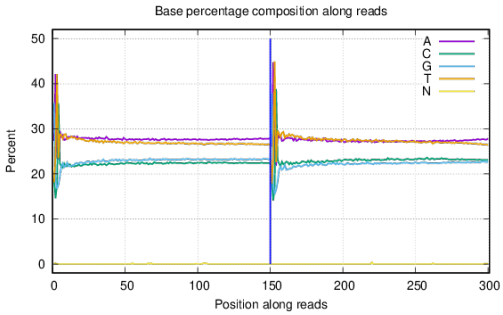


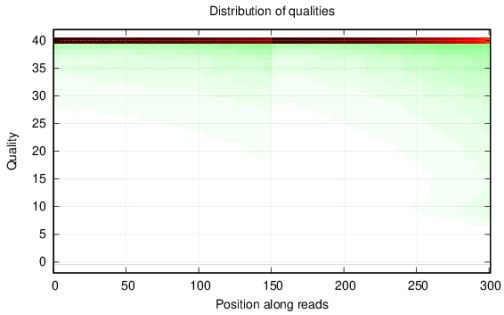
Quality control of sample Liv_hic_wt_01





Quality control of sample Liv_hic_wt_03





4 Help Document

The original image data is transferred into sequence data via base calling, which is defined as raw data or raw reads and saved as FASTQ file. Each entry in a FASTQ files consists of 4 lines:

1. A sequence identifier with information about the sequencing run and the cluster. The exact contents of this line vary by based on the BCL to FASTQ conversion software used.
2. The sequence (the base calls; A, C, T, G and N).
3. A separator, which is simply a plus (+) sign.
4. The base call quality scores. These are Phred +33 encoded, using ASCII characters to represent the numerical quality scores.

Here is an example of a single entry in a FASTQ file:

@V300029029L1C001R0010000210/1
GCGACCCCAGGTCAGTCGGGACTACCCGCTGAAGTCGGAGGCCAAGCGGT
+
FFFCFFFFFFFFFDFEFFFFEFEF0FFFFEFFFFFFFEFFFFFECGFFFF

The relationship between DNBseq sequencer sequencing error rate and the sequencing quality value is shown in the following formula. Specifically, if the sequencing error rate is denoted as "E", DNBseq sequencer base quality value is denoted as "sQ", the relationship is as follows:

$$sQ = -10\log_{10} E$$

Sequencing error rate	Sequencing quality value	Character of Phred +33 quality system
5%	13	.
1%	20	5
0.1%	30	?